The Grand Design Challenge for Music GenAI

Nicholas J. Bryan
Head of Music AI
Adobe Research


AI-generated music will outnumber all human recorded music ever made in a few years
— 100+ years of recorded history surpassed.

"Copyright Act of 1976 requires all eligible work to be authored in the first instance by a human being”

***Thaler v. Perlmutter*, No. 23-5233 (D.C. Cir. March 18, 2025)**
***Thaler v. Perlmutter*, No. 25-449 (SCOTUS, cert. denied March 2, 2026)**

~All music AI-generated

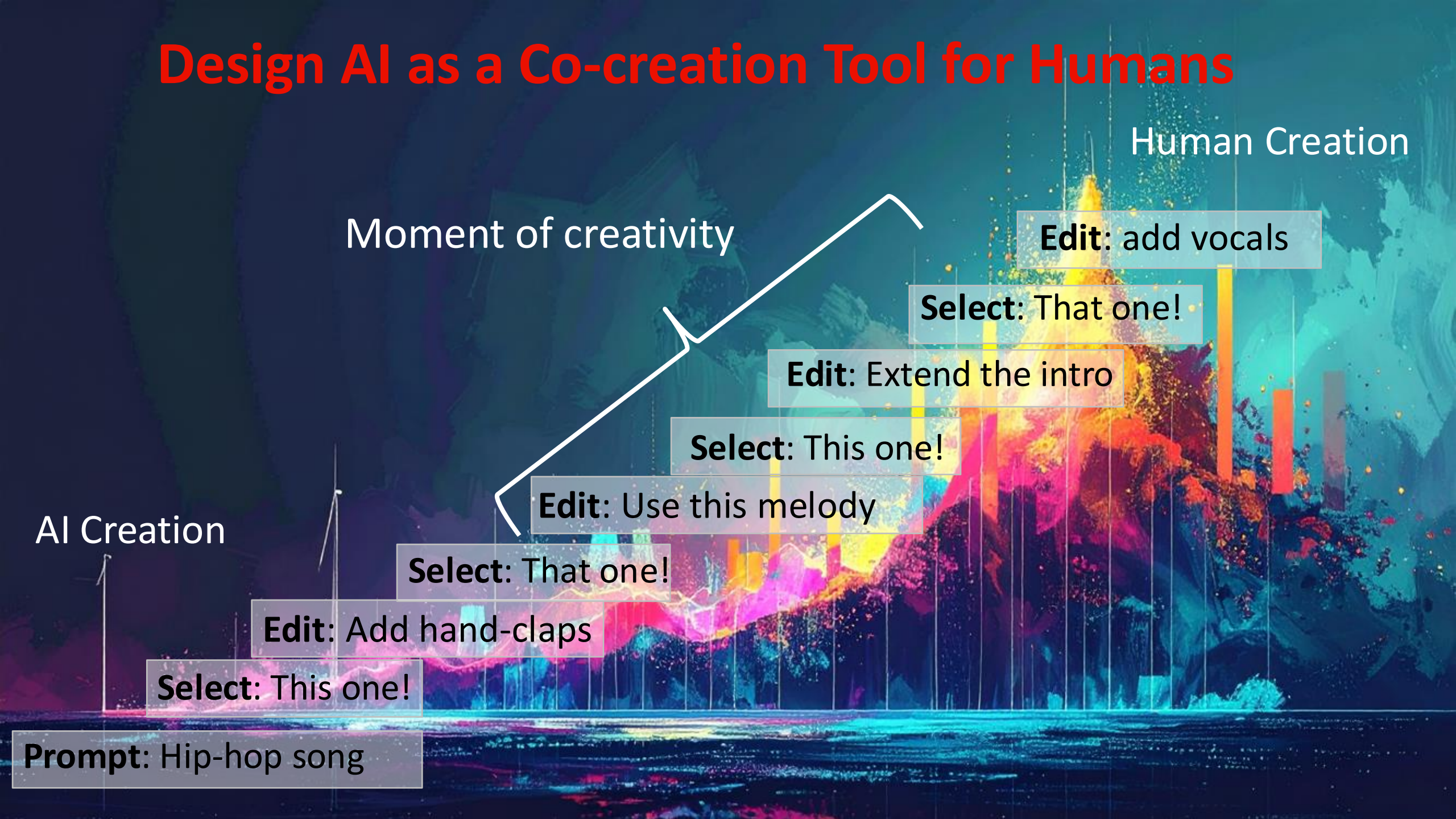
~No copyrights for AI music

"I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes."

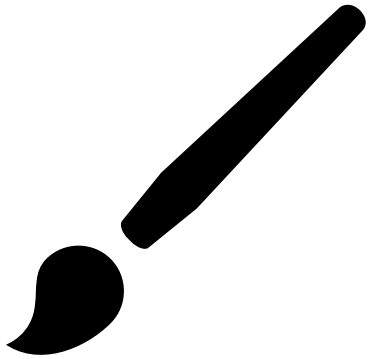
Author and videogame enthusiast **Joanna Maciejewska** nails it (although bathroom cleaning goes ahead of laundry and dishes)

serve

Design AI as a Co-creation Tool for Humans



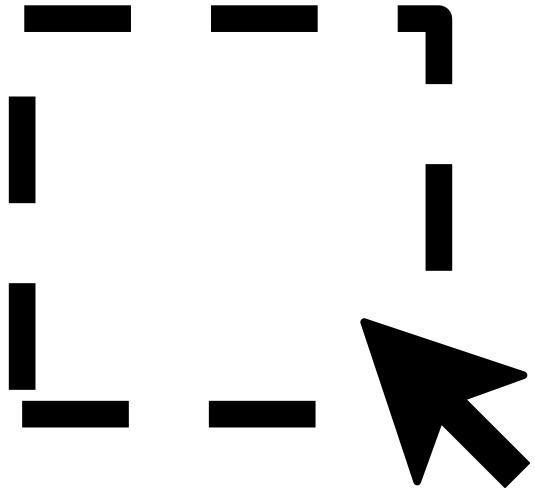
A Design Language for Music Co-Creation



Control

- Genre
- Mood
- Instruments
- Keywords
- …

- Melody
- Intensity
- Rhythm
- Structure
- Multi-track



Editing

- Looping
- Extension
- Replacement
- Tempo
- Key

- Add instrument
- Remove instrument
- Separate
- …

Recent Work

Recent Work

arXiv:2311.07069v1 [cs.SD] 13 Nov 2023

MUSIC CONTROLNET: Multiple Time-varying Controls for Music Generation

Shih-Lun Wu[1,2*], Chris Donahue[1], Shinji Watanabe[1], and Nicholas J. Bryan[2]

[1] School of Computer Science, Carnegie Mellon University [2] Adobe Research

***Abstract*—Text-to-music generation models are now capable of generating high-quality music audio in broad styles. However, text control is primarily suitable for the manipulation of *global* musical attributes like genre, mood, and tempo, and is less suitable for precise control over *time-varying* attributes such as the positions of beats in time or the changing dynamics of the music. We propose Music ControlNet, a diffusion-based music generation model that offers multiple precise, time-varying controls over generated audio. To imbue text-to-music models with time-varying control, we propose an approach analogous to pixel-wise control of the image-domain ControlNet method. Specifically, we extract controls from training audio yielding paired data, and fine-tune a diffusion-based conditional generative model over audio spectrograms given melody, dynamics, and rhythm controls. While the image-domain Uni-ControlNet method already allows generation with any subset of controls, we devise a new masking strategy at training to allow creators to input controls that are only partially specified in time. We evaluate both on controls extracted from audio and controls we expect creators to provide, demonstrating that we can generate realistic music that corresponds to control inputs in both settings. While few comparable music generation models exist, we benchmark against MusicGen, a recent model that accepts text and melody input, and show that our model generates music that is 49% more faithful to input melodies despite having 35x fewer parameters, training on 11x less data, and enabling two additional forms of time-varying control. Sound examples can be found at https://MusicControlNet.github.io/web/.**

***Index Terms*—music generation, controllable generative modeling, diffusion**

I. INTRODUCTION

One of the pillars of musical expression is the communication of high-level ideas and emotions through precise manipulation of lower-level attributes like notes, dynamics, and rhythms. Recently, there has been an explosion of interest in training text-to-music generative models that allow creators to directly convert high-level intent (expressed as text) into music audio [1, 2, 3, 4, 5]. These models suggest an exciting new paradigm of musical expression wherein creators can instantaneously generate realistic music without the need to write a melody, specify meter and rhythm, or orchestrate instruments. However, while dramatically more efficient, this new paradigm ignores more conventional forms of musical expression rooted in the manipulation of lower-level attributes, limiting the ability to express precise musical intent or leverage models in existing creative workflows.

There are two primary obstacles for adding precise control to text-based music generation methods. Firstly, relative to symbolic music representations like scores, text is a cumbersome interface for conveying precise musical attributes that vary over time. Verbose and mundane text descriptions may be needed to precisely represent even the first note of a musical score e.g., "the song starts at 80 beats per minute with a quarter note on middle C played mezzo-forte on the saxophone". The second obstacle is an empirical one—text-to-music models tend to faithfully interpret *global* stylistic attributes (e.g., genre and mood) from text, but struggle to interpret text descriptions of precise musical attributes (e.g., notes or rhythms). This is perhaps a consequence of the relative scarcity of precise descriptions in the training data.

A potential solution to the musical imprecision of natural language is the incorporation of *time-varying* controls into music generation. For example, one body of work looks at synthesizing music audio from time-varying symbolic music representations like MIDI [6, 7], however this approach offers a particularly strict form of control requiring users to compose entire pieces of music beforehand. Such approaches are more similar to typical music composition processes and do not take full advantage of recent text-to-music methods. Another body of work on musical style transfer [8, 9, 10, 11, 12, 13] seeks to transform recordings from one *style* (e.g., genre, musical ensemble, or mood) to another while preserving the underlying composition content. However, a majority of these approaches require training an individual model per style, as opposed to the flexibility of using text to control style in a single model.

In this work, we propose Music ControlNet, a diffusion-based music generation model that offers multiple time-varying controls over the melody, dynamics, and rhythm of generated audio, in addition to global text-based style control as shown in Fig. 1. To incorporate such time-varying controls, we adapt recent work on image generation with spatial control, namely, ControlNet [14] and Uni-ControlNet [15] to enable musical controls that are *composable* (i.e., can generate music corresponding to any subset of controls) and further allow creators to only *partially specify* each of the controls both for convenience and to direct our model to musically *improvise* in remaining time spans of the generation. To overcome the aforementioned scarcity of precise, ground-truth control inputs, following [5, 16], we extract useful control signals directly from music during training. We evaluate our method on two different categories of control signals: (1) *extracted* control signals that come from example songs, which are similar to those seen during training, and (2) *created* control

*Work done during Shih-Lun's internship at Adobe Research. Correspondence should be addressed to Shih-Lun Wu and Nicholas J. Bryan at shihlunw@cs.cmu.edu and njb@ieee.org, respectively.

Music ControlNet

arXiv:2602.09891v1 [cs.SD] 10 Feb 2026

STEMPHONIC: ALL-AT-ONCE FLEXIBLE MULTI-STEM MUSIC GENERATION

Shih-Lun Wu[♯♭*] *Ge Zhu*[♭] *Juan-Pablo Caceres*[♭] *Cheng-Zhi Anna Huang*[♯] *Nicholas J. Bryan*[♭]

[♯] MIT CSAIL [♭] Adobe Research

ABSTRACT

Music stem generation, the task of producing musically-synchronized and isolated instrument audio clips, offers the potential of greater user control and better alignment with musician workflows compared to conventional text-to-music models. Existing stem generation approaches, however, either rely on fixed architectures that output a predefined set of stems in parallel, or generate only one stem at a time, resulting in slow sequential inference despite flexibility in stem combination. We propose STEMPHONIC, a diffusion-/flow-based framework that overcomes this trade-off and generates a variable set of synchronized stems in one inference pass. During training, we treat each stem as a batch element, group synchronized stems in a batch, and apply a shared noise latent to each group. At inference-time, we use a shared initial noise latent and stem-specific text inputs to generate synchronized multi-stem outputs in one pass. We further expand our approach to enable one-pass conditional multi-stem generation and stem-wise activity controls to empower users to iteratively generate and orchestrate the temporal layering of a mix. We benchmark our results on multiple open-source stem evaluation sets and show that STEMPHONIC produces higher-quality outputs while accelerating the full mix generation process by 25–50%. Demos at: https://stemphonic-demo.vercel.app.

***Index Terms*— music audio generation, stem generation, conditional stem generation, variable stem combinations, diffusion, flow.**

1. INTRODUCTION

Text-to-audio music generation models are now able to produce realistic sounding music from simple text inputs [1–5]. They lower the barrier for music creation, enabling anyone to explore and express their creativity, but typically generate *fully-mixed* multi-instrument outputs that are difficult to edit and cannot easily be reused as components in new compositions [6, 7]. To empower creators beyond text prompting, numerous control and editing methods have been proposed, including fine-grained temporal controls [8–10], music inpainting [10–12], as well as music stem generation [13–16]. A music *stem* is a recording of one or more instruments that collectively serve as a distinct layer in a mix, e.g., *drums* for rhythmic foundation, and *basses* for low-pitch progressions. Generating stems enables creators to edit each stem separately and experiment with different mixing/mastering techniques, enhancing their creative control [17, 18].

Existing stem generation methods can largely be classified into (*1*) models with a *parallelized* architecture, and (*2*) individual-stem models that require *sequential* generation of stems. Parallelized models [13–16] generate *multiple* coherent stems in a *single pass*, but handle limited, coarse-grained stem types (e.g., only basses, drums, vocals, others) that must be fixed in advance and built into the architecture. Individual-stem models [19–23], on the other hand, allow

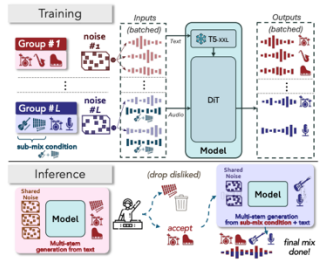


Fig. 1. Our STEMPHONIC framework for flexible multi-stem music generation. (Top) At training, each group of synchronized stems share the same noise latent. (Bottom) At inference, we use a shared initial noise to generate variable multi-stem outputs in one pass. We also enable conditional stem generation and stem-wise activity controls.

flexible open-vocabulary stem generations through text prompting or other conditioning mechanisms, and can often condition on existing audio to iteratively generate new accompanying stems and create a mix with an arbitrary number of stems. Yet, these models generate stems *one at a time*, leading to slower full inference processes.

To unify the strengths of both paradigms and alleviate their drawbacks, we propose STEMPHONIC, a latent diffusion/flow-based [24, 25] framework capable of generating a variable set of musically-synchronized stems in one inference pass, as shown in Fig. 1. We introduce two techniques applied during training. First, we treat each stem as a batch element and group musically-synchronized stems in a batch (Sec. 3.1). Then, we assign a single shared noise latent to each group (Sec. 3.2). At inference, we use a shared initial noise and different stem-specific text inputs to generate variably many synchronized stem outputs in one pass. We further expand our approach to conditional multi-stem generation, and add stem-wise activity controls (Sec. 3.3) that enable creators to iteratively generate and orchestrate the temporal layering of a mix. In our experiments, we ablate our stem grouping and noise sharing core techniques, and verify the conditional generation and stem-wise activity control capabilities. We find STEMPHONIC capable of generating higher-quality multi-stem mixes, while accelerating the full inference process by 25–50%, compared to the existing individual-stem iterative workflow.

Our contributions can be summarized as follows:

- A latent diffusion/flow-based framework to generate a variable number of synchronized stems efficiently in one inference pass,

* Work done while an intern at Adobe Research.

Stemphonic

arXiv:2603.11042v1 [cs.CV] 11 Mar 2026

V2M-ZERO: Zero-Pair Time-Aligned Video-to-Music Generation

Yan-Bo Lin[1*] Jonah Casebeer[2] Long Mai[2]
Aniruddha Mahapatra[2] Gedas Bertasius[1] Nicholas J. Bryan[2]

[1]UNC Chapel Hill [2]Adobe Research

Abstract. Generating music that temporally aligns with video events is challenging for existing text-to-music models, which lack fine-grained temporal control. We introduce V2M-ZERO, a zero-pair video-to-music generation approach that outputs time-aligned music for video. Our method is motivated by a key observation: temporal synchronization requires matching when and how much change occurs, not what changes. While musical and visual events differ semantically, they exhibit shared temporal structure that can be captured independently within each modality. We capture this structure through event curves computed from intra-modal similarity using pretrained music and video encoders. By measuring temporal change within each modality independently, these curves provide comparable representations across modalities. This enables a simple training strategy: fine-tune a text-to-music model on music-event curves, then substitute video-event curves at inference without cross-modal training or paired data. Across **OES-Pub**, **MovieGenBench-Music**, and **AIST++**, V2M-ZERO achieves substantial gains over paired-data baselines: **5–21%** higher audio quality, **13–15%** better semantic alignment, **21–52%** improved temporal synchronization, and **28%** higher beat alignment on dance videos. We find similar results via a large crowd-source subjective listening test. Overall, our results validate that temporal alignment through within-modality features, rather than paired cross-modal supervision, is effective for video-to-music generation. Results are available at https://genjib.github.io/v2m_zero/

1 Introduction

Generative music is growing in popularity among creators from online influencers on social media platforms (e.g., YouTube, Instagram, TikTok) to professionals in film, gaming, and advertising. Such content creators seek music that both complements their video content and supports fast and flexible control over style and pacing. While recent text-to-music (T2M) methods [1, 15, 21, 24, 40, 57, 63, 73, 81, 101, 110] enable automatic music generation from textual prompts, their outputs are not designed to follow the temporal dynamics of a target video. As

* Work done during an internship at Adobe Research.

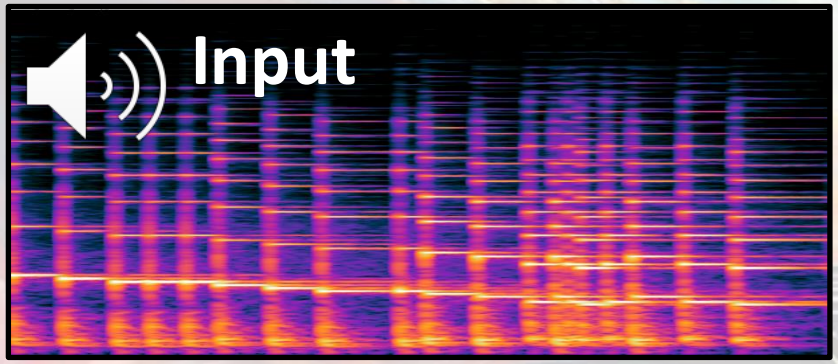
V2M-Zero

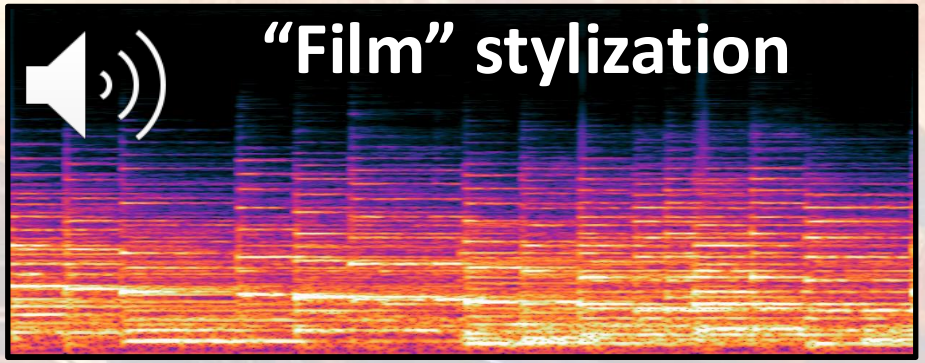
Music ControlNet:

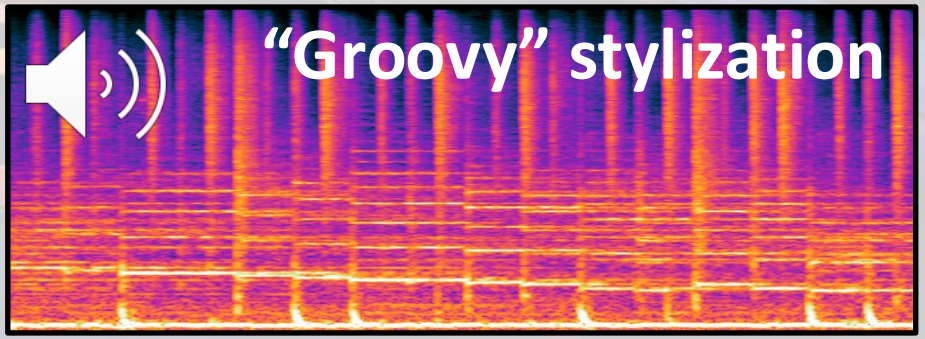
Multiple Time-varying Controls for Music Generation

S.L. Wu, C. Donahue, S. Watanabe, N. J. Bryan “Music ControlNet: Multiple Time-varying Controls for Music Generation,” IEEE TASLP 2024.

Melody Control



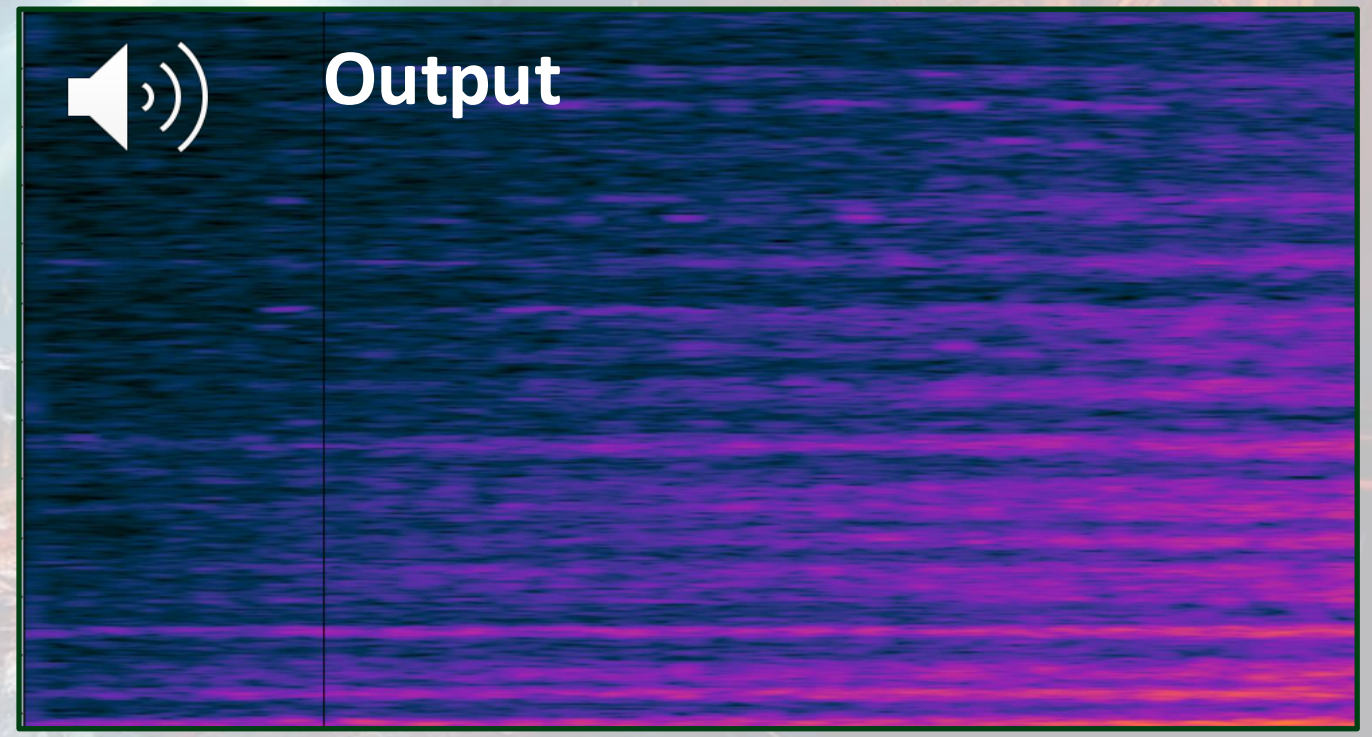




S.L. Wu, C. Donahue, S. Watanabe, N. J. Bryan “Music ControlNet: Multiple Time-varying Controls for Music Generation,” IEEE TASLP 2024.

Intensity Control

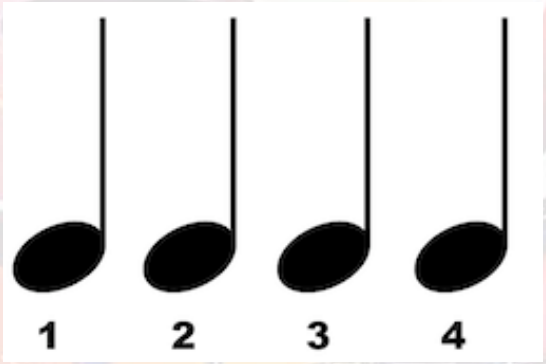
Input intensity

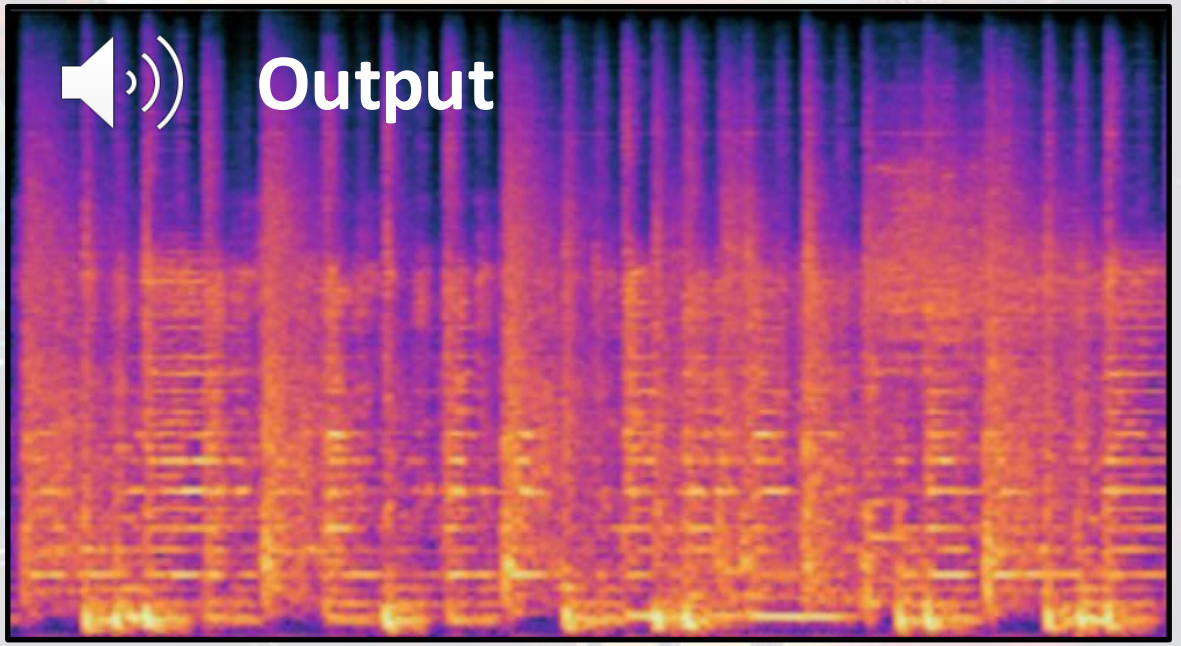


S.L. Wu, C. Donahue, S. Watanabe, N. J. Bryan “Music ControlNet: Multiple Time-varying Controls for Music Generation,” IEEE TASLP 2024.

Rhythm Control

Input click





S.L. Wu, C. Donahue, S. Watanabe, N. J. Bryan “Music ControlNet: Multiple Time-varying Controls for Music Generation,” IEEE TASLP 2024.

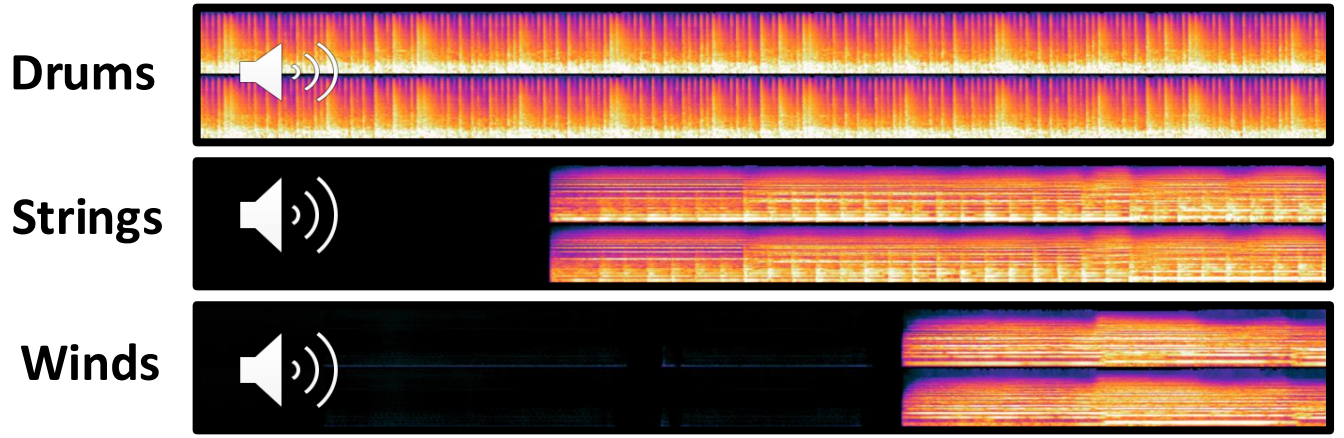
Stemphonic:

All-at-Once Flexible Multi-stem Music Generation

S.L. Wu, G. Zhu, J.P. Caceres, C-Z. A. Huang, N. J. Bryan. “Stemphonic: All-at-Once Flexible Multi-Stem Music Generation,” IEEE ICASSP 2026.

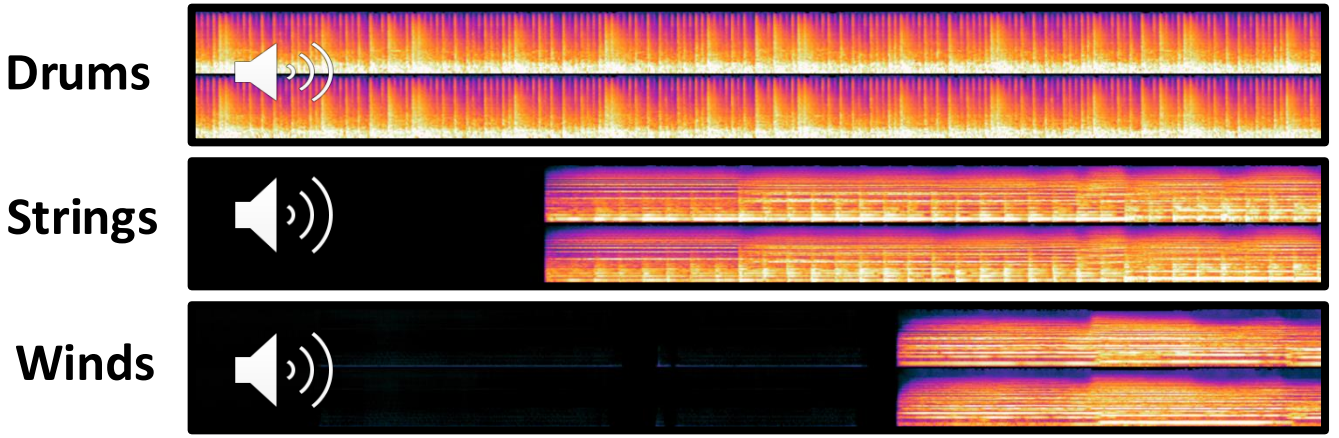
Generate Multi-track

- An intense epic orchestral piece with **drum**, **string**, and **winds** tracks

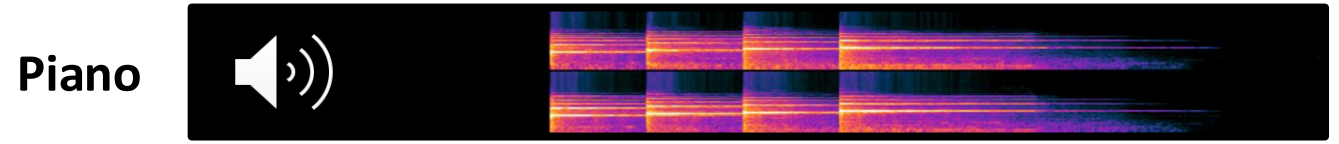


Add Track (and Remove Track)

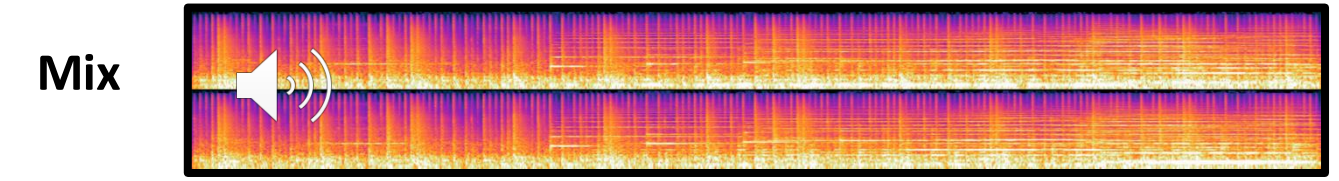
- An intense epic orchestral piece with **drum**, **string**, and **winds** tracks



- Add **piano**



- Final mix



V2M-Zero:

Zero-Pair Time-Aligned Video-to-Music Generation

Y-B Lin, J. Casebeer, L. Mai, A. Mahapatra, G. Bertasius , N. J. Bryan. “V2M-Zero: Zero-Pair Time-Aligned Video-to-Music Generation,” arXiv 2026.

How?

Diffusion & Flow Models

Fixed Forward Diffusion Process

$\mathbf{x}_0$  $\mathbf{x}_T$

Generative Reverse Diffusion Process (Learned)

ϵ_θ

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *NeurIPS* 2020

Sampling

Fixed Forward Diffusion Process

$\mathbf{x}_0$  $\mathbf{x}_T$

Generative Reverse Diffusion Process (Learned)

Const. from forward process

Randomness

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\right) + \sigma_t \mathbf{z}$$

New Estimate

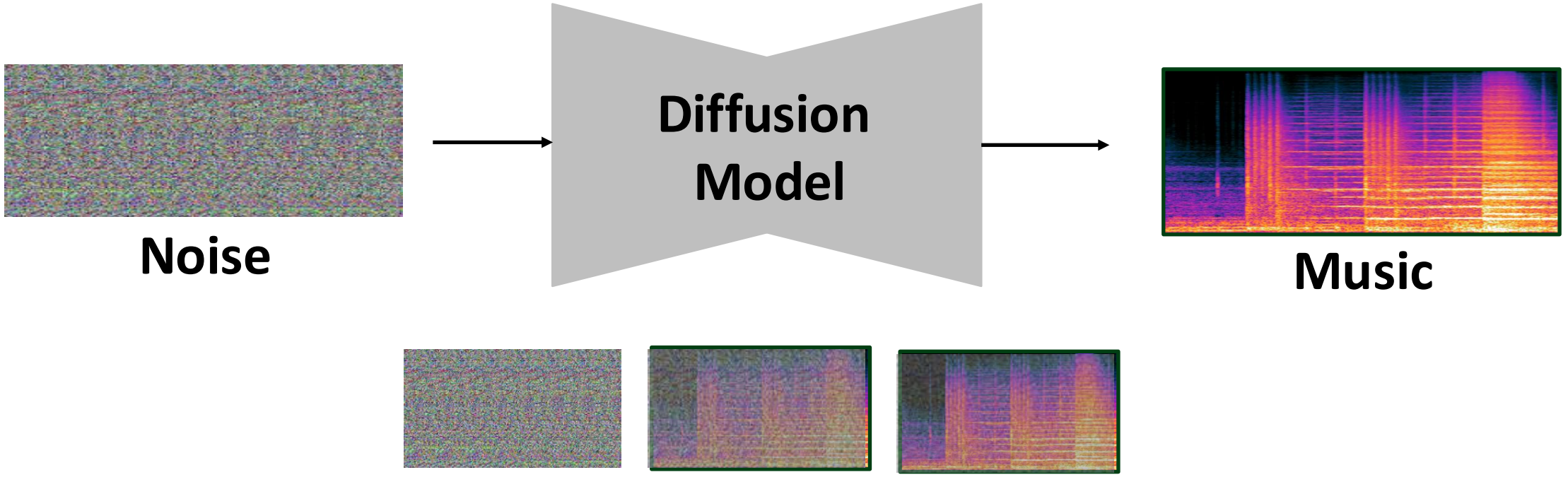
Old Estimate

Controls

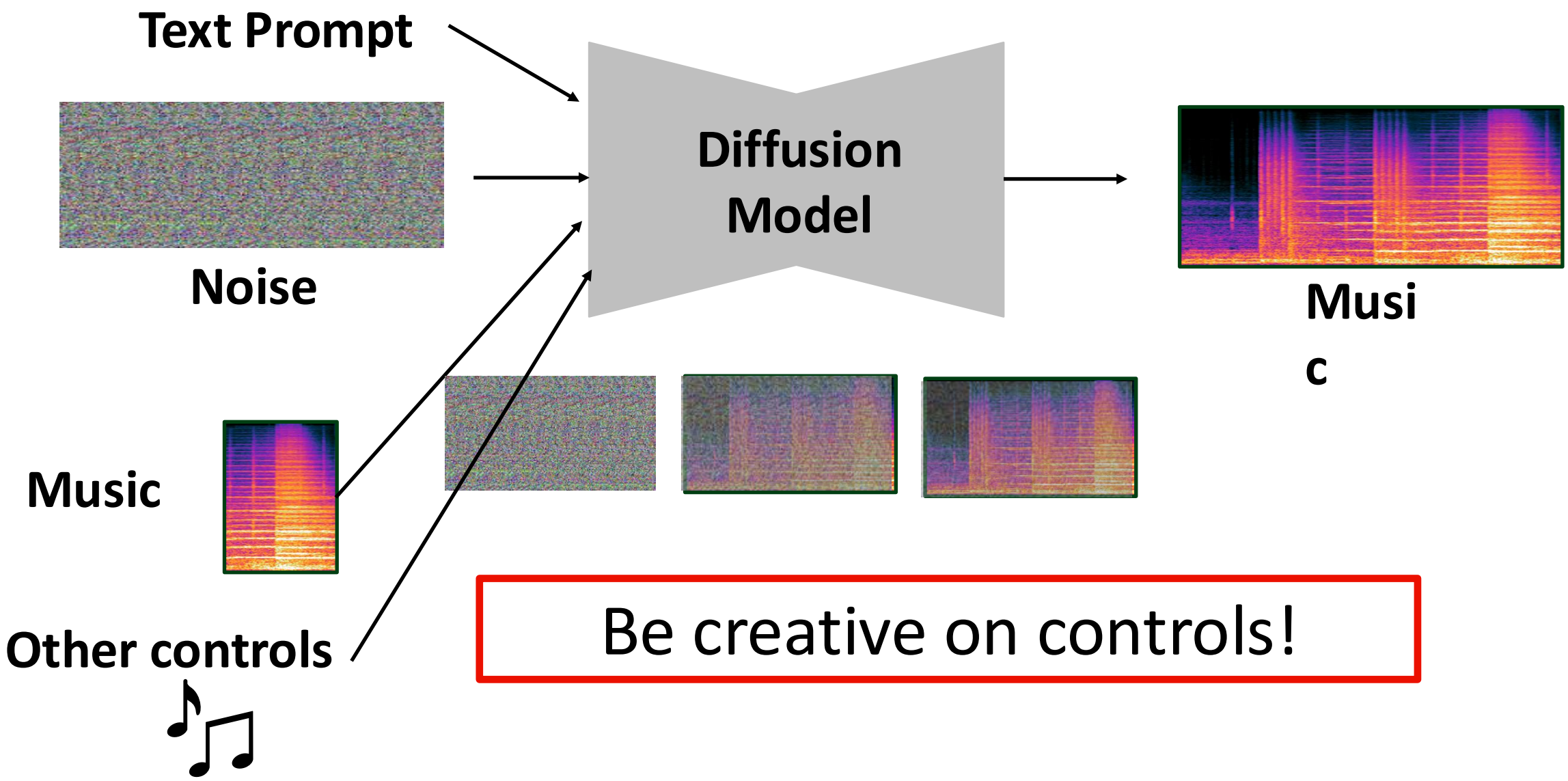
Learned Network

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *NeurIPS* 2020

Diffusion Model



Diffusion Model w/Controls



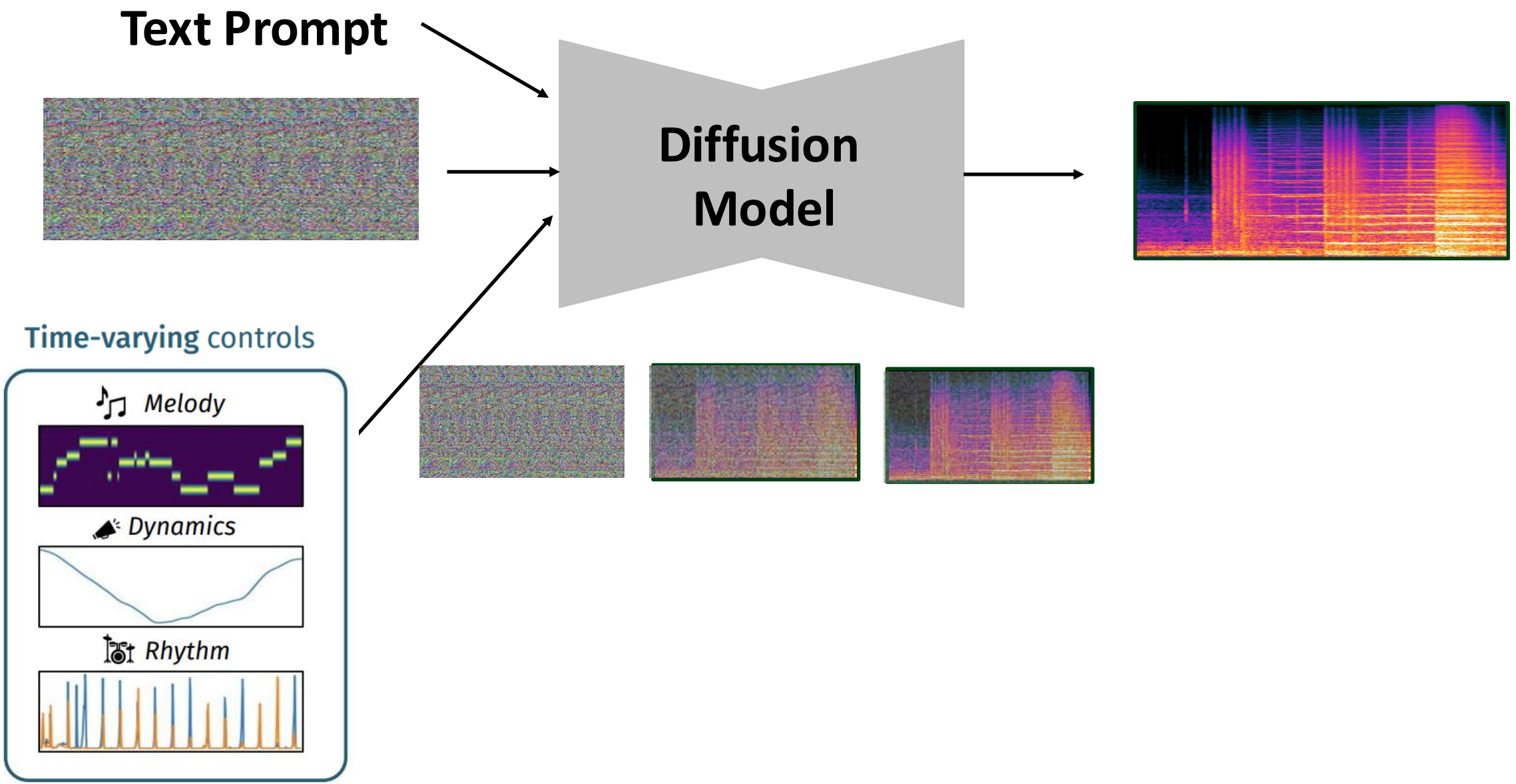
Three Strategies for Adding Controls

Pretraining

Fine-tuning

Inference-time

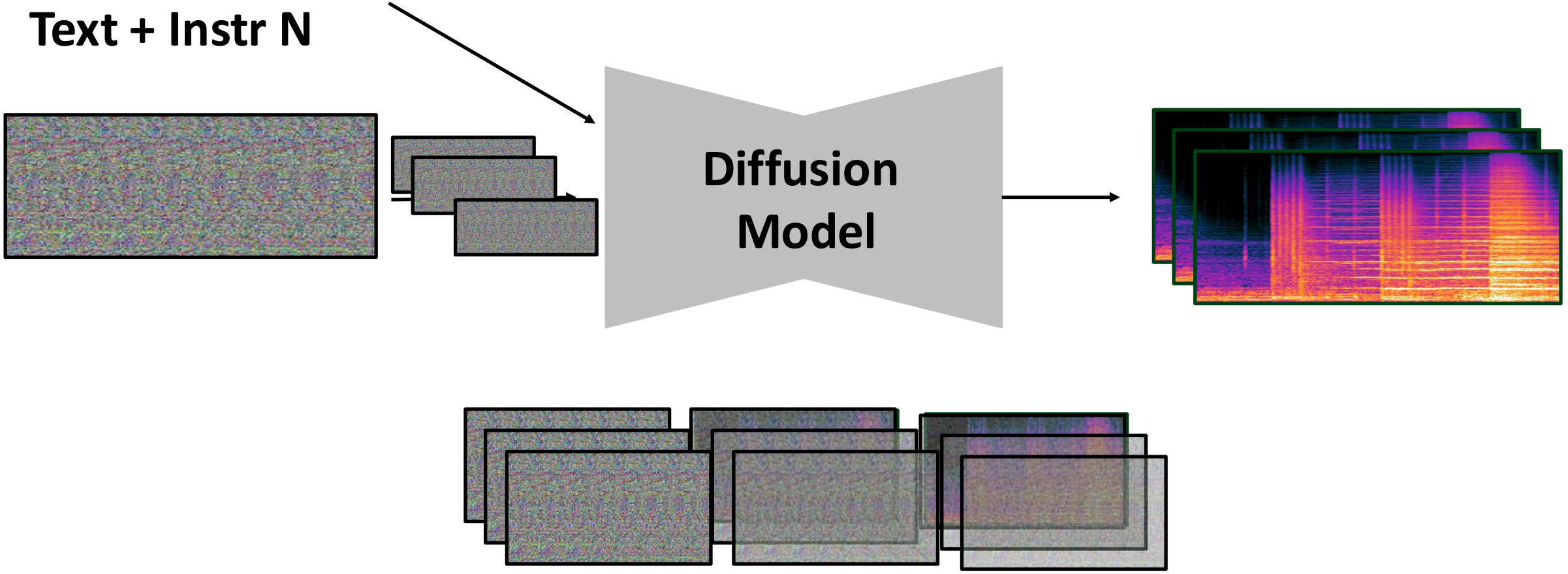
Music ControlNet: Multiple Time-Varying Controls for Music Generation



Stemphonic: All-At-Once Flexible Multi-Stem Generation

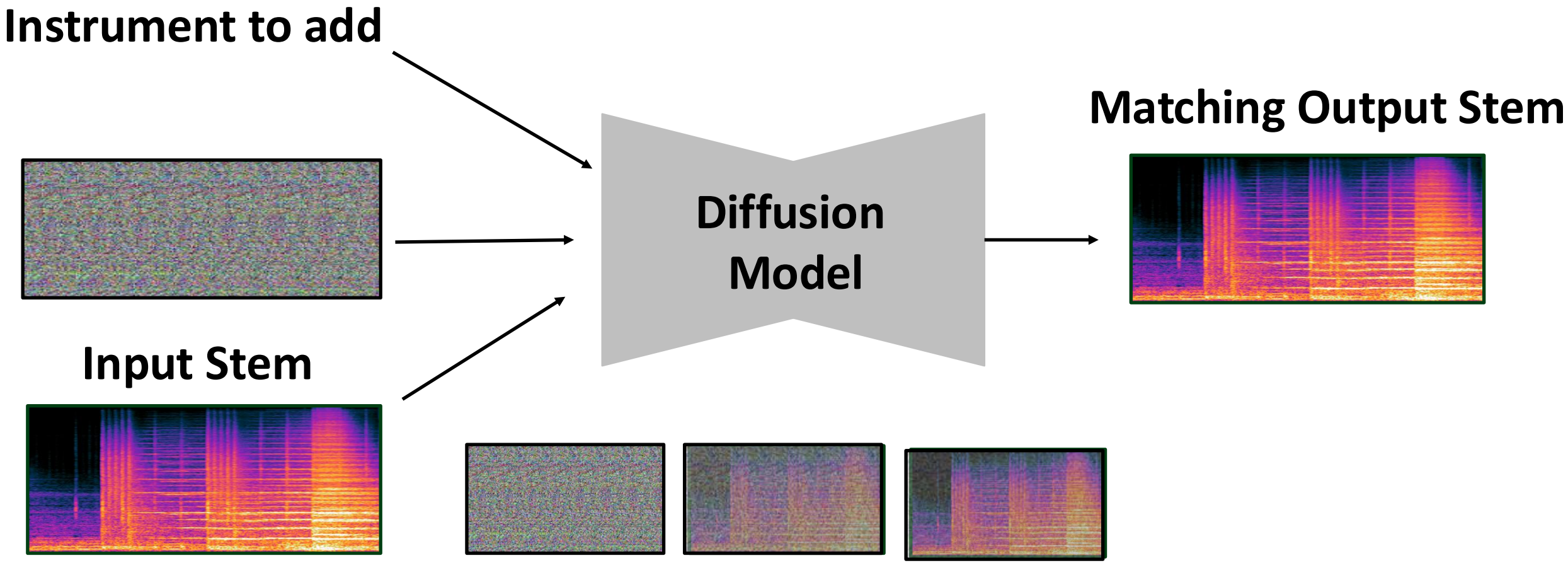
Multi-track generation

Text + Instr 1
Text + Instr 2
…
Text + Instr N



Stemphonic: All-At-Once Flexible Multi-Stem Generation

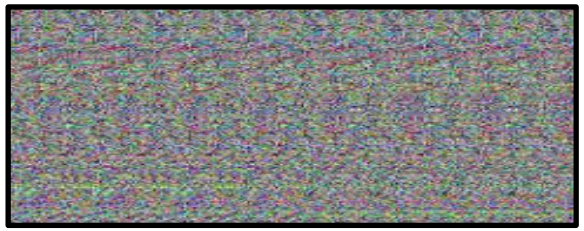
Accompaniment generation



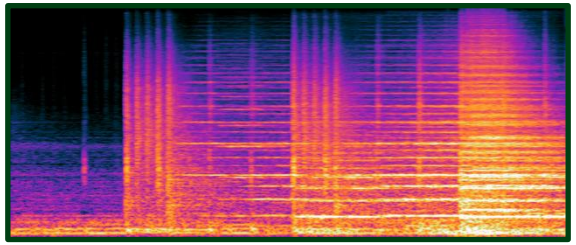
Stemphonic: All-At-Once Flexible Multi-Stem Generation

Instrument Removal

Instrument to remove

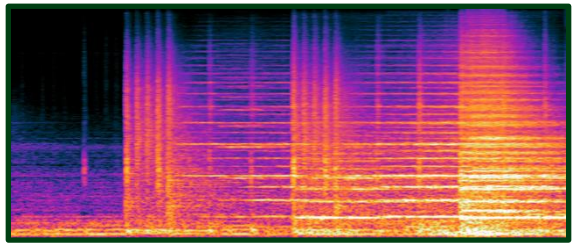


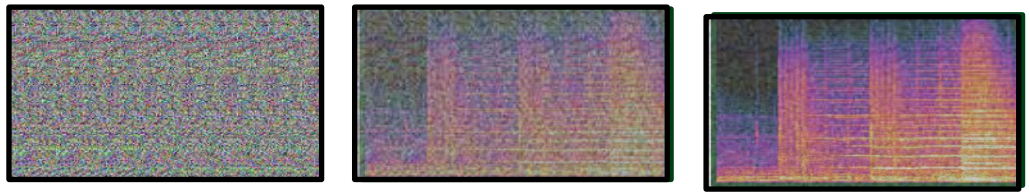
Input Mixture



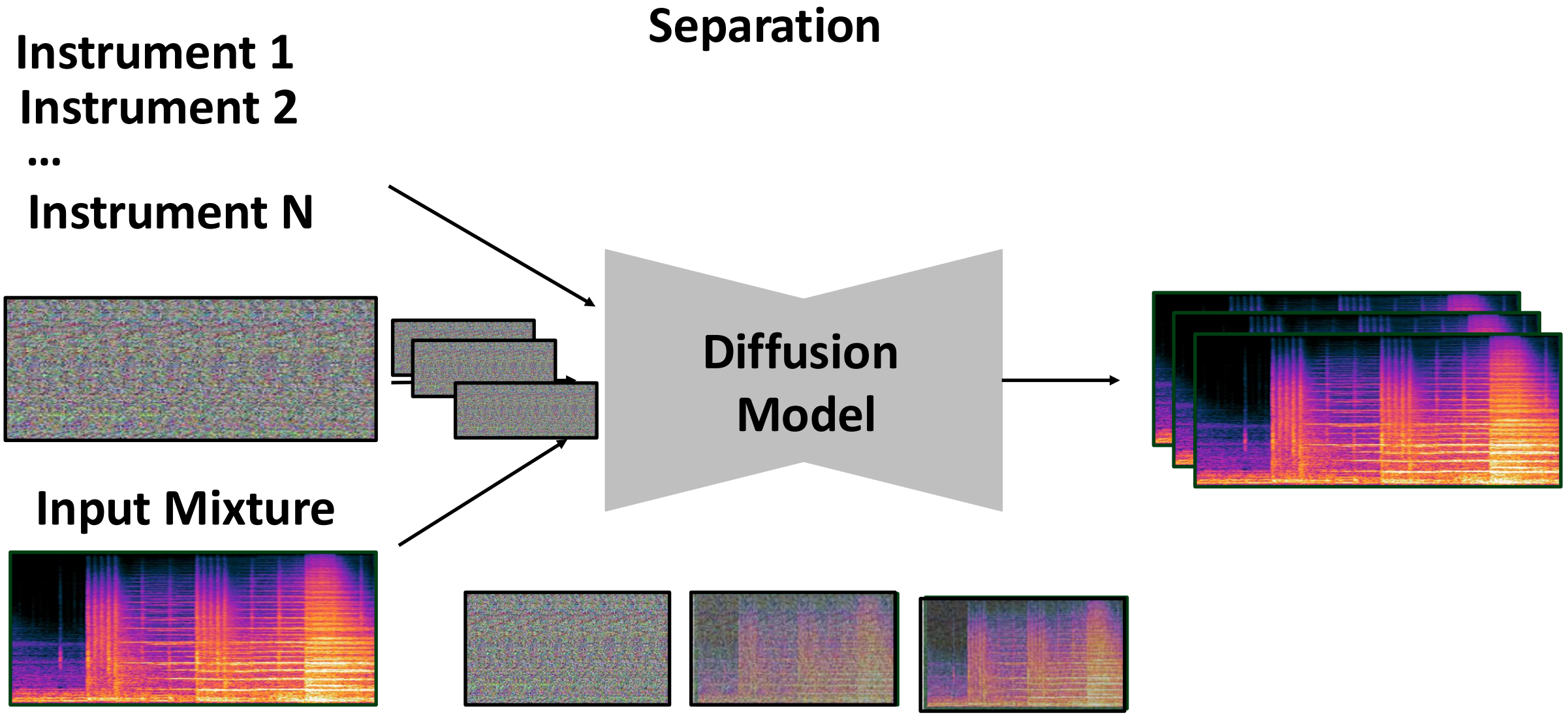
Diffusion Model

Mix minus Instr

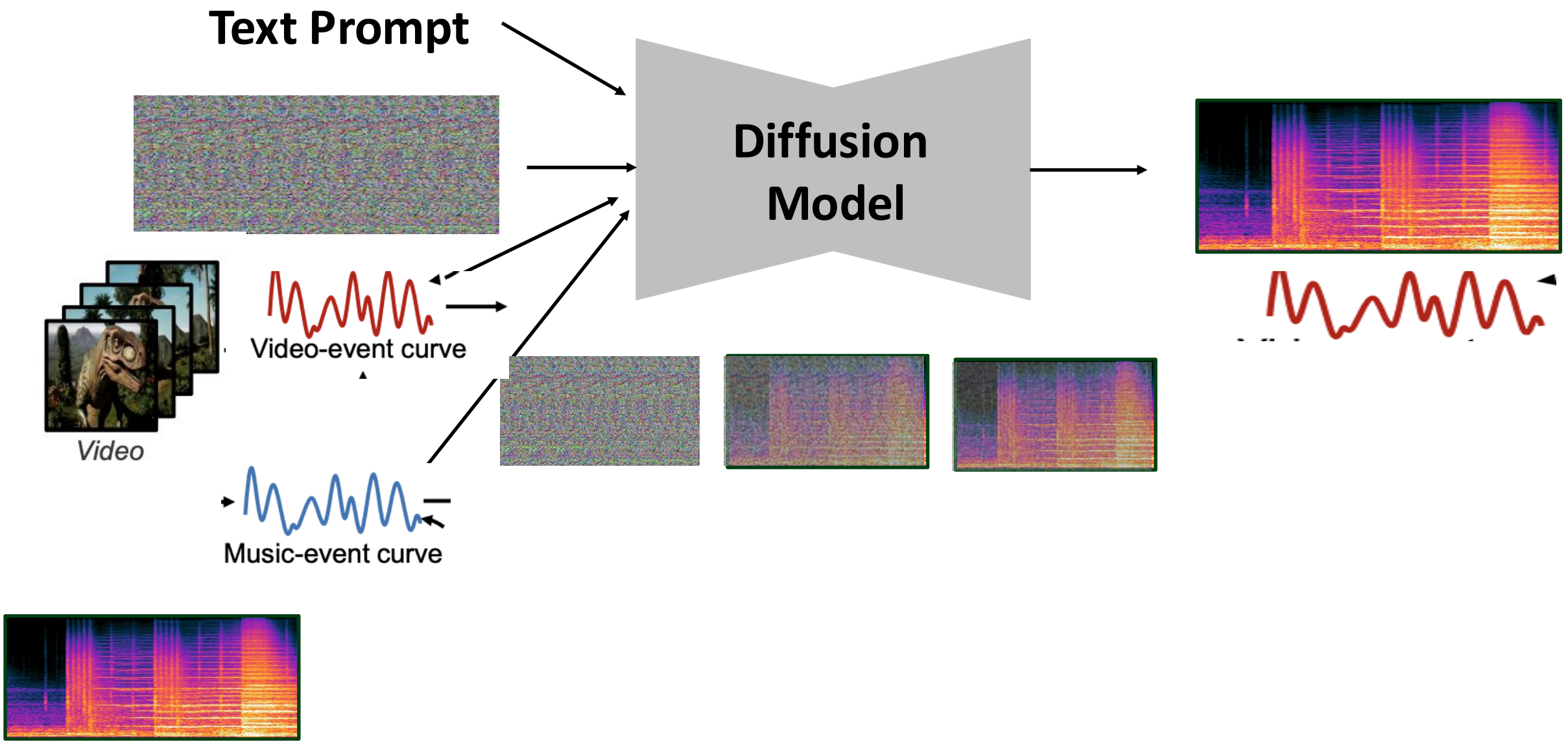




Stemphonic: All-At-Once Flexible Multi-Stem Generation



V2M-Zero: Zero-Pair Time-Aligned Video-to-Music Generation



Make it Fast

Recent Work

Presto! DISTILLING STEPS AND LAYERS FOR ACCELERATING MUSIC GENERATION

Zachary Novack*
UC – San Diego

Ge Zhu & Jonah Casebeer
Adobe Research

Julian McAuley & Taylor Berg-Kirkpatrick
UC – San Diego

Nicholas J. Bryan
Adobe Research

arXiv:2410.05167v2 [cs.SD] 16 Apr 2025

ABSTRACT

Despite advances in diffusion-based text-to-music (TTM) methods, efficient, high-quality generation remains a challenge. We introduce ***Presto!***, an approach to inference acceleration for score-based diffusion transformers via reducing both sampling steps and cost per step. To reduce steps, we develop a new score-based distribution matching distillation (DMD) method for the EDM-family of diffusion models, the first GAN-based distillation method for TTM. To reduce the cost per step, we develop a simple, but powerful improvement to a recent layer distillation method that improves learning via better preserving hidden state variance. Finally, we combine our step and layer distillation methods together for a dual-faceted approach. We evaluate our step and layer distillation methods independently and show each yield best-in-class performance. Our combined distillation method can generate high-quality outputs with improved diversity, accelerating our base model by 10-18x (230/435ms latency for 32 second mono/stereo 44.1kHz, 15x faster than the comparable SOTA model) — the fastest TTM to our knowledge.

1 INTRODUCTION

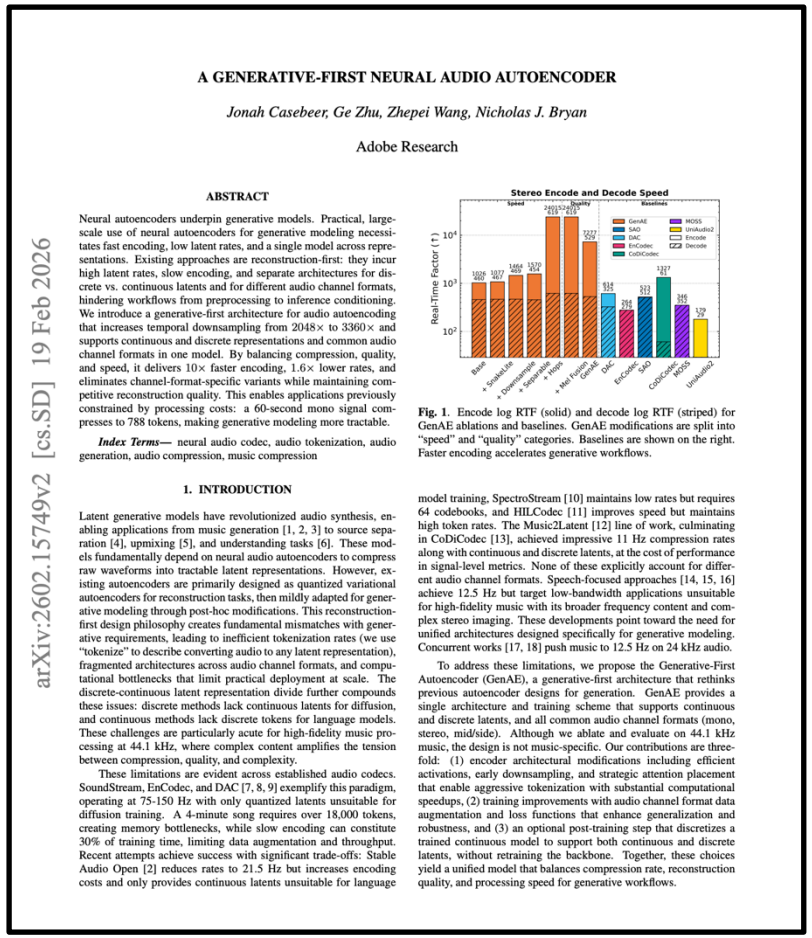
We have seen a renaissance of audio-domain generative media (Chen et al., 2024; Agostinelli et al., 2023; Liu et al., 2023; Copet et al., 2023), with increasing capabilities for both Text-to-Audio (TTA) and Text-to-Music (TTM) generation. This work has been driven in-part by audio-domain *diffusion models* (Song et al., 2020; Ho et al., 2020; Song et al., 2021), enabling considerably better audio modeling than generative adversarial network (GAN) or variational autoencoder (VAE) methods (Dhariwal & Nichol, 2021). Diffusion models, however, suffer from long inference times due to their iterative denoising process, requiring a substantial number of function evaluations (NFE) during inference (i.e. sampling) and resulting in ≈5-20 seconds at best for non-batched ≈32s outputs.

Accelerating diffusion inference typically focuses on *step distillation*, i.e. the process of reducing the *number* of sampling steps by distilling the diffusion model into a few-step generator. Methods include consistency-based (Salimans & Ho, 2022; Song et al., 2023; Kim et al., 2023) and adversarial (Sauer et al., 2023; Yin et al., 2023; 2024; Kang et al., 2024) approaches. Others have also investigated *layer-distillation* (Ma et al., 2024; Wimbauer et al., 2024; Moon et al., 2024), which draws from transformer early exiting (Hou et al., 2020; Schuster et al., 2021) by dropping interior layers to reduce the *cost* per sampling step for image generation. For TTA/TTM models, however, distillation techniques have only been applied to shorter or lower-quality audio (Bai et al., 2024; Novack et al., 2024a), necessitate ≈10 steps (vs. 1-4 step image methods) to match base quality (Saito et al., 2024), and have not successfully used layer or GAN-based distillation methods.

We present **Presto**[1], a dual-faceted distillation approach to inference acceleration for score-based diffusion transformers via reducing the number of sampling steps and the cost per step. **Presto** includes three distillation methods: (1) **Presto-S**, a new distribution matching distillation algorithm for *score-based*, EDM-style diffusion models (see Fig. 1) leveraging GAN-based step distillation

*Work done while an intern at Adobe. Correspondence to znovack@ucsd.edu and njb@ieee.org.
[1]*Presto* is the common musical term denoting fast music from 168-200 beats per minute.

Presto!
(fast inference)



A GENERATIVE-FIRST NEURAL AUDIO AUTOENCODER

Jonah Casebeer, Ge Zhu, Zhepei Wang, Nicholas J. Bryan

Adobe Research

arXiv:2602.15749v2 [cs.SD] 19 Feb 2026

ABSTRACT

Neural autoencoders underpin generative models. Practical, large-scale use of neural autoencoders for generative modeling necessitates fast encoding, low latent rates, and a single model across representations. Existing approaches are reconstruction-first: they incur high latent rates, slow encoding, and separate architectures for discrete vs. continuous latents and for different audio channel formats, hindering workflows from preprocessing to inference conditioning. We introduce a generative-first architecture for audio autoencoding that increases temporal downsampling from 2048× to 3360× and supports continuous and discrete representations and common audio channel formats in one model. By balancing compression, quality, and speed, it delivers 10× faster encoding, 1.6× lower rates, and eliminates channel-format-specific variants while maintaining competitive reconstruction quality. This enables applications previously constrained by processing costs: a 60-second mono signal compresses to 788 tokens, making generative modeling more tractable.

Index Terms— neural audio codec, audio tokenization, audio generation, audio compression, music compression

Fig. 1. Encode log RTF (solid) and decode log RTF (striped) for GenAE ablations and baselines. GenAE modifications are split into "speed" and "quality" categories. Baselines are shown on the right. Faster encoding accelerates generative workflows.

1. INTRODUCTION

Latent generative models have revolutionized audio synthesis, enabling applications from music generation [1, 2, 3] to source separation [4], upmixing [5], and understanding tasks [6]. These models fundamentally depend on neural audio autoencoders to compress raw waveforms into tractable latent representations. However, existing autoencoders are primarily designed as quantized variational autoencoders for reconstruction tasks, then mildly adapted for generative modeling through post-hoc modifications. This reconstruction-first design philosophy creates fundamental mismatches with generative requirements, leading to inefficient tokenization rates (we use "tokenize" to describe converting audio to any latent representation), fragmented architectures across audio channel formats, and computational bottlenecks that limit practical deployment at scale. The discrete-continuous latent representation divide further compounds these issues: discrete methods lack continuous latents for diffusion, and continuous methods lack discrete tokens for language models. These challenges are particularly acute for high-fidelity music processing at 44.1 kHz, where complex content amplifies the tension between compression, quality, and complexity.

These limitations are evident across established audio codecs. SoundStream, EnCodec, and DAC [7, 8, 9] exemplify this paradigm, operating at 75-150 Hz with only quantized latents unsuitable for diffusion training. A 4-minute song requires over 18,000 tokens, creating memory bottlenecks, while slow encoding can constitute 30% of training time, limiting data augmentation and throughput. Recent attempts achieve success with significant trade-offs: Stable Audio Open [2] reduces rates to 21.5 Hz but increases encoding costs and only provides continuous latents unsuitable for language model training, SpectroStream [10] maintains low rates but requires 64 codebooks, and HILCodec [11] improves speed but maintains high token rates. The Music2Latent [12] line of work, culminating in CoDiCodec [13], achieved impressive 11 Hz compression rates along with continuous and discrete latents, at the cost of performance in signal-level metrics. None of these explicitly account for different audio channel formats. Speech-focused approaches [14, 15, 16] achieve 12.5 Hz but target low-bandwidth applications unsuitable for high-fidelity music with its broader frequency content and complex stereo imaging. These developments point toward the need for unified architectures designed specifically for generative modeling. Concurrent works [17, 18] push music to 12.5 Hz on 24 kHz audio.

To address these limitations, we propose the Generative-First Autoencoder (GenAE), a generative-first architecture that rethinks previous autoencoder designs for generation. GenAE provides a single architecture and training scheme that supports continuous and discrete latents, and all common audio channel formats (mono, stereo, mid/side). Although we ablate and evaluate on 44.1 kHz music, the design is not music-specific. Our contributions are threefold: (1) encoder architectural modifications including efficient activations, early downsampling, and strategic attention placement that enable aggressive tokenization with substantial computational speedups, (2) training improvements with audio channel format data augmentation and loss functions that enhance generalization and robustness, and (3) an optional post-training step that discretizes a trained continuous model to support both continuous and discrete latents, without retraining the backbone. Together, these choices yield a unified model that balances compression rate, reconstruction quality, and processing speed for generative workflows.

GenAE
(fast training)

Presto! Distilling Steps and Layers for Accelerating Music Generation

Zachary Novack[♯♭*], Ge Zhu[♭], Jonah Casebeer[♭]

Julian McAuley[♯], Taylor Berg-Kirkpatrick[♯], Nicholas J. Bryan[♭]

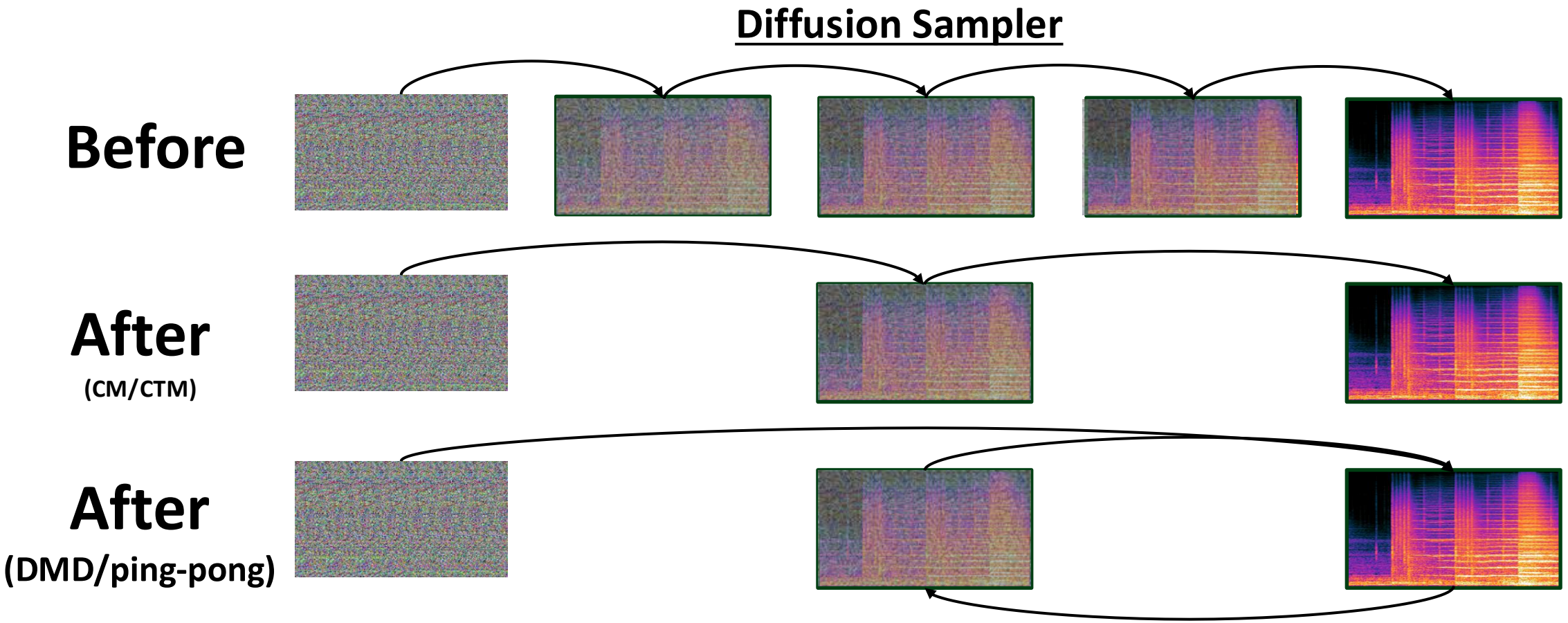
[♯]UCSD

[♭]Adobe Research

* Work done during an internship at Adobe Research.

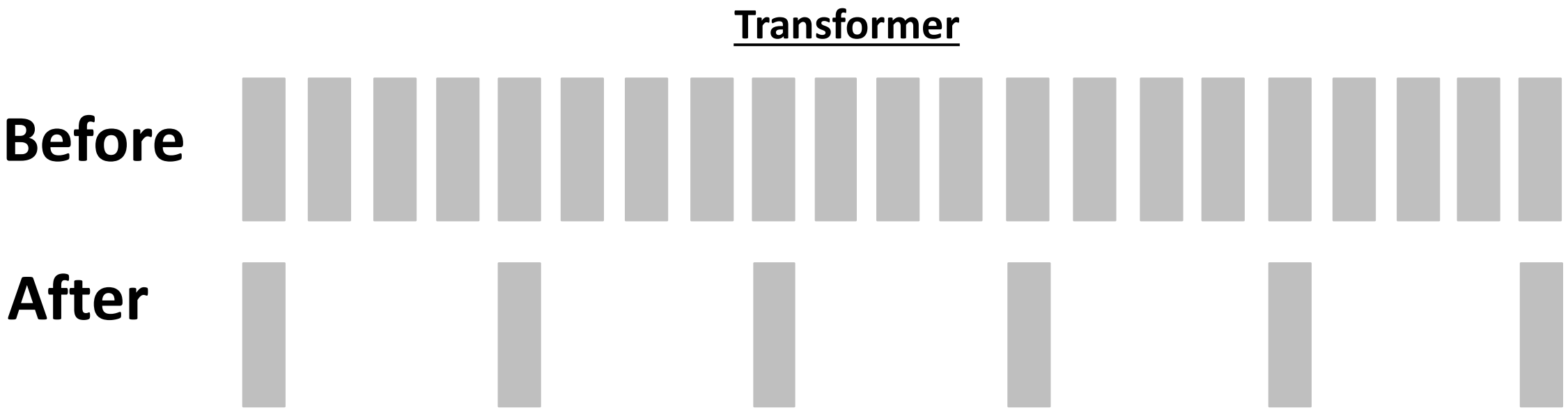
(Listen with headphones)

Distilling Steps

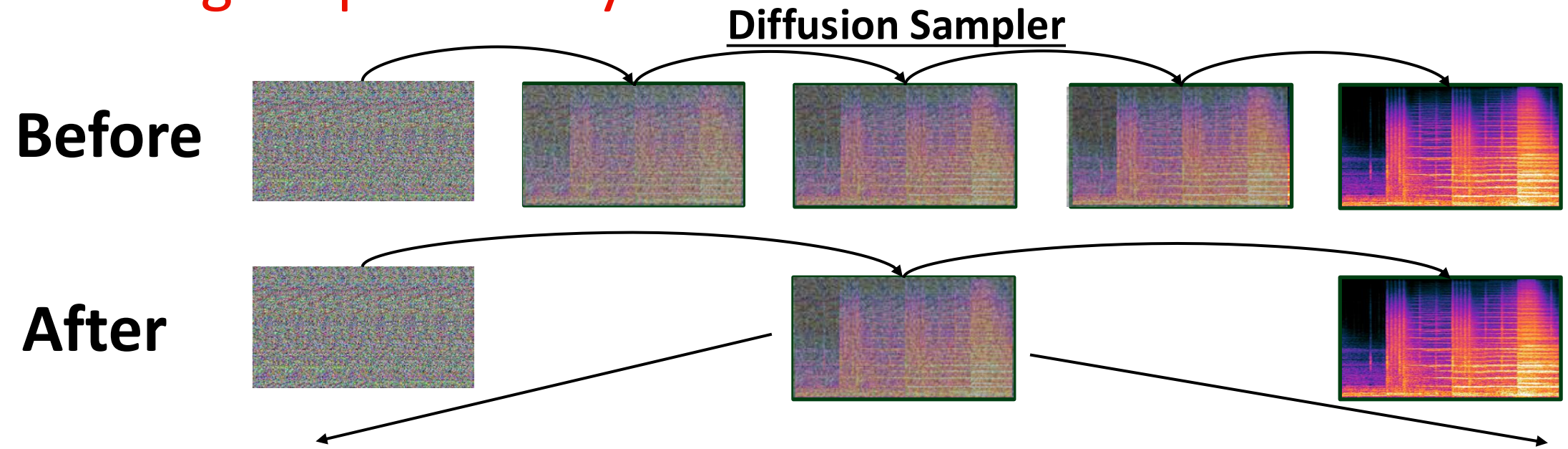


Z. Novack, G. Zhu, J. Casebeer, J. McAuley, T. Berg-Kirkpatrick, N. J Bryan. “Presto! Distilling Steps and Layers For Accelerating Music Generation,” ICLR 2025

Distilling Layers (of a Transformer)



Distilling Steps and Layers



GenAE

A Generative-First Neural Audio Autoencoder

Jonah Casebeer, Ge Zhu, Zhepei Wang, Nicholas J. Bryan

Adobe Research
Accepted to IEEE ICASSP 2026

We present GenAE, a neural audio

References

- S-L. Wu, C. Donahue, S. Watanabe, and N. J. Bryan, “Music ControlNet: Multiple Time-varying Controls for Music Generation.” IEEE Trans. on Audio, Speech, and Lang. Proc. (TASLP), 2023.
- S-L. Wu, G. Zhu, J-P. Caceres, C-Z. Huang, N. J. Bryan. “Stemphonic: All-at-once Flexible Multi-stem Music Generation.” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026.
- Y-B. Lin, J. Casebeer, L. Mai, A. Mahapatra, G. Bertasius, N. J. Bryan. “V2M-Zero: Zero-Shot Time-Aligned Video-to-Music Generation.” arXiv, 2025.
- Z. Novack, G. Zhu, J. Casebeer, J. McAulay, T. Berg-Kirkpatrick, N. J. Bryan, “Presto! Distilling Steps and Layers for Accelerating Music Generation.” *International Conference on Learning Representations (ICLR),* 2025.
- J. Casebeer, G. Zhu, Zhepei Wang, N. J. Bryan. “A Generative-first Neural Audio Autoencoder.” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026.

Conclusion

- Grand Design Challenge for Music GenAI!
- Design AI as co-creation tool for human
- Three examples of adding controls to pretrained music models
  - Music ControlNet – Melody, intensity, and rhythm control
  - Stemphonic – Multi-track generation, accompaniment generation (removal, and separation)
  - V2M-Zero – Time-aligned Video-to-Music Generation without paired data
  - Fine-tuning controls is super powerful!
- Two examples of making models fast (inference + training)